(2)

Are All Linear Paired Comparison Models Equivalent?

Hal Stern
Department of Statistics
Harvard University

Technical Report No. ONR-C-5

September, 1990

Are All Linear Paired Comparison Models Equivalent?

Hal Stern

Department of Statistics

Harvard University

Cambridge, MA 02138 U.S.A.

ABSTRACT

Previous authors (Jackson and Fleckenstein 1957, Mosteller 1958, Noether 1960) have found that different models of paired comparisons data lead to similar fits. This phenomenon is examined by means of a set of paired comparison models, based on gamma random variables, that includes the frequently applied Bradley-Terry and Thurstone-Mosteller models. A theoretical result provides a natural ordering of the models in the gamma family on the basis of their composition rules. Analysis of several sports data sets indicates that all of the paired comparison models in the family provide adequate, and almost identical, fits to the data. Simulations are used to further explore this result. Although not all approaches to paired comparisons experiments are covered by this discussion, the evidence is strong that for samples of the size usually encountered in practice all linear paired comparison models are virtually equivalent.

Abbreviated Title: Comparing Paired Comparison Models

Keywords: Bradley-Terry Model, Thurstone-Mosteller Model

1. INTRODUCTION

In a paired comparisons experiment, k objects are compared in blocks of size two. Each comparison of two objects has two possible outcomes: either i is preferred to j or j is preferred to i. Successive comparisons of a pair of objects are assumed to be independent. In addition, comparisons of distinct pairs of objects are assumed to be independent of each other. This eliminates the notion of a single judge who compares each of the $\binom{k}{2}$ distinct pairs, as the comparisons in this case would almost certainly not be independent. A variety of models exist for the analysis of data from paired comparisons experiments, including the Bradley-Terry model (Bradley and Terry 1952) and the Thurstone-Mosteller model (Thurstone 1927, Mosteller 1951). Jackson and Fleckenstein (1957) and Mosteller (1958) illustrate that these two models, as well as several others, provide similar fits to a data set.

A family of paired comparison models based on gamma random variables (Stern 1990) provides a framework for further consideration of the similarity of paired comparison models. The gamma paired comparison models are a subset of the class of linear models (David 1988) that includes the Bradley-Terry and Thurstone-Mosteller models. The probability that i is preferred to j in a gamma paired comparison model with shape parameter r is equal to the probability that one gamma random variable with shape parameter r is smaller than a second gamma random variable, independent of the first, with the same shape parameter but different scale parameter. This model is appropriate, for example, if we compare the waiting time until r events occur in each of two independent Poisson processes with different rates. The Bradley-Terry model is obtained by choosing $r = 1$ and the Thurstone-Mosteller model is obtained as $r \to \infty$. In these cases, equivalence to the usual stochastic utility model is obtained by considering a logarithmic transformation of the gamma random variables (Stern 1990).

Evidence from three different sources indicates that, for typical sample sizes, the choice of a particular paired comparison model from among the set of gamma models seems to have a small effect on the results obtained. In this paper, analysis of several sports data sets indicates that almost identical fits are obtained by several models. Close consideration of the case with $k = 3$ objects provides some information about the source of the problem and provides an estimate of the sample size required to distinguish between paired comparison models. Finally, some simulations generalize the results to larger experiments.

In the next three sections, a variety of paired comparison models are discussed. The evidence concerning the question in the title of the paper is presented in Section 5.

2. PAIRED COMPARISON MODELS

The natural parameter in a paired comparisons experiment is p_{ij}, the probability that i is preferred to j. Probability models for paired comparisons experiments attempt to provide a concise description of the preference probabilities p_{ij}, $i \neq j$. Sophisticated models have been developed to account for the possibility of ties, covariates and order effects. For the purposes of this discussion, only paired comparison models that ignore ties, order effects and covariates are considered. By assumption, the preference probability p_{ij} remains constant throughout the experiment. The saturated model for a paired comparisons experiment with k objects associates a parameter p_{ij} with the pair of objects i and j, thus using $k(k-1)/2$ parameters. A more parsimonious model assigns a parameter λ_i to each object and takes $p_{ij} = P(\lambda_i, \lambda_j)$ for some function $P(\cdot, \cdot)$. This type of model uses only k parameters. The Bradley-Terry and Thurstone-Mosteller models are examples of

this type. These models are now considered in more detail, leading to a family of paired comparison models used throughout this study.

The Bradley-Terry probability model assumes the probability that i is preferred to j can be written as

$$p_{ij}^{(BT)} = \frac{\lambda_i}{\lambda_i + \lambda_j}.$$

Over time this expression has been derived in many ways including a derivation based on Luce's (1959) Choice Axiom and one based on maximum entropy (Joe 1987). Two motivations that are central to this paper are the gamma random variable motivation (Stern 1990) and the linear model derivation (David 1988, Latta 1979). Throughout the paper, paired comparisons experiments are discussed using the terminology of a sports competition because the data in Section 5 is of this form. Suppose that team i scores points according to a Poisson process with rate λ_i and team j scores points according to a Poisson process with rate λ_j. Furthermore, suppose the two Poisson processes are independent. The waiting time for a point to be scored in either process is an exponential random variable, or equivalently, a gamma random variable with shape parameter 1. Then, the probability that team i scores one point before team j is the probability that $X_i \sim \Gamma(1, \lambda_i)$ (X_i a gamma random variable with shape parameter 1 and scale parameter λ_i) is less than $X_j \sim \Gamma(1, \lambda_j)$ for independent random variables X_i, X_j. This probability is the Bradley-Terry preference probability (Bradley and Terry 1952). Holman and Marley derived the Bradley-Terry model in terms of exponential random variables (equivalent to the above formulation) in 1962 (see Luce and Suppes 1965).

Other gamma paired comparison models are obtained by comparing gamma random variables with shape parameters other than one. The point scoring motivation suggests models with integer-valued shape parameter, but gamma random

variables are defined for any shape parameter r greater than zero. Suppose that $G_\lambda(r)$ is a stochastic process with independent increments having the gamma distribution, so that $G_\lambda(r_2) - G_\lambda(r_1)$ has the gamma distribution with shape parameter $r_2 - r_1$ and scale parameter λ. Thus far, $G_\lambda(r)$ has been interpreted for integer r as the waiting time for r points to be scored. However, the progress of two gamma stochastic processes $G_{\lambda_i}(r)$ and $G_{\lambda_j}(r)$ can be compared for any value of $r > 0$ suggesting the possibility of gamma paired comparison models with non-integer shape parameters.

For the gamma paired comparison model with shape parameter r, the preference probability is given by

$$p_{ij}^{(r)} = \Pr(X_i < X_j) = \int_0^\infty \int_0^{x_j} \frac{\lambda_i^r x_i^{r-1} \exp(-\lambda_i x_i)}{\Gamma(r)} \frac{\lambda_j^r x_j^{r-1} \exp(-\lambda_j x_j)}{\Gamma(r)} dx_i dx_j$$

$$= \int_0^\infty \int_0^{\frac{\lambda_i}{\lambda_j} z_j} \frac{z_i^{r-1} \exp(-z_i)\ z_j^{r-1} \exp(-z_j)}{\Gamma(r)\Gamma(r)} dz_i dz_j = g_r\left(\frac{\lambda_i}{\lambda_j}\right). \qquad (1)$$

The final notation indicates that this probability depends only on the ratio of the scale parameters of the gamma random variables. Since the probability is unchanged if each λ_i is multiplied by the same constant, $\sum \lambda_i = 1$ is adopted as a convention. By reversing the roles of i and j, the natural relationship $g_r(\lambda_i/\lambda_j) = 1 - g_r(\lambda_j/\lambda_i)$ is obtained. The preference probability is increasing in the ratio λ_i/λ_j, and for $\lambda_i > \lambda_j$, $p_{ij}^{(r)}$ is increasing in r. The first of these results indicates that the probability that the process with the higher rate is the first to achieve r points increases as the difference between the rates of the two processes becomes larger. This is easy to verify by inspection of the expression (1). The second result implies that, if i scores points faster than j, then comparing the processes after they have evolved for a long time favors process i. If we take $\gamma = \lambda_i/\lambda_j$, this can be demonstrated by

examining $\frac{\partial p_{ij}^{(r)}}{\partial r}$ and $\frac{\partial^2 p_{ij}^{(r)}}{\partial r \partial \gamma}$ as functions of γ and r. The first derivative is equal to zero at $\gamma = 1$ and tends to zero as $\gamma \to \infty$ for any r. The mixed second derivative is positive at $\gamma = 1$ for any r, so the first derivative is positive for γ slightly larger than one. The second derivative remains positive until some critical value after which it is always negative. Given this second derivative behavior, $\frac{\partial p_{ij}^{(r)}}{\partial r}$ must be positive for all $\lambda_i > \lambda_j$. It follows that when $\lambda_i < \lambda_j$, $p_{ij}^{(r)}$ decreases as r increases.

Following David (1988), a set of preference probabilities $p_{ij}, i \neq j$ are said to satisfy a linear model if there exist real numbers $v_1, \ldots, v_k$ such that $p_{ij} = H(v_i - v_j)$ for $H(\cdot)$ monotone increasing from $H(-\infty) = 0$ to $H(\infty) = 1$ with $H(x) = 1 - H(-x)$. The function $H(\cdot)$ is the cumulative distribution function (c.d.f.) of a random variable symmetric around zero and is called the defining distribution of the linear model. The parameters v_i measure the positions of the k teams on a linear scale. A linear model with defining distribution H is called a convolution type linear model if H can be derived as the distribution of the difference between two independent random variables with common c.d.f. $F(\cdot)$ and different location parameters. In this case F is called the sensation distribution of the convolution type linear model. The Bradley-Terry model is obtained by taking $v_i = \ln \lambda_i$ and $v_j = \ln \lambda_j$, with $H(x) = (1 + \exp(-x))^{-1}$, the c.d.f. of the logistic distribution (Bradley 1953), and $F(x) = \exp(-e^{-x})$, the c.d.f. of the extreme value distribution (Davidson 1969).

It turns out that, for any r, the gamma paired comparison model can be expressed as a convolution type linear model where the density of the sensation distribution is

$$f_r(x) = \frac{1}{\Gamma(r)} e^{-r(x - v_i)} e^{-\exp(-(x - v_i))}$$

with $v_i = \ln \lambda_i$, and the density of the corresponding defining distribution is

$$h_r(x) = \frac{\Gamma(2r)}{\Gamma(r)\Gamma(r)} \frac{e^{-rx}}{(1+e^{-x})^{2r}}.$$

Integrating $h_r(\cdot)$ and evaluating the result at $x = \ln \lambda_i - \ln \lambda_j$ leads to the preference probability given in expression (1). As $r \to \infty$, a gamma random variable with shape parameter r tends to a normally distributed random variable. Thus the gamma model with r large is similar to a convolution type linear paired comparison model with Gaussian sensation distribution, and therefore Gaussian defining distribution. The Gaussian linear model is described by Thurstone (1927) and refined by Mosteller (1951). For small values of r, the gamma model is similar to the convolution type linear model whose sensation distribution is the ordinary exponential distribution with a location parameter (Mosteller 1958, Noether 1960). Formal statements describing the limiting behavior of the gamma model for small or large r can be found in Stern (1987, 1990).

For the remainder of this article, discussion is focused on gamma paired comparison models, or equivalently the subset of convolution type linear models that they represent. This is a particularly interesting family because it includes the most popular approaches to paired comparisons experiments in a single family, indexed by the single parameter r. Naturally, there are linear models that are not convolution type linear models (the uniform model considered by Smith (1956), Mosteller (1958), Noether (1960)) and other convolution type linear models (for example, those with the Student's t-distribution or the Cauchy distribution as the defining distribution) that are not considered here. Thus, any answer to the question posed by the title of the article is incomplete. Nonetheless, the evidence indicates that, within the class of linear models, all models are essentially equivalent. In order to further discuss the empirical evidence, we consider Latta's (1979) partial ordering

of paired comparison models.

3. COMPOSITION RULES AND A PARTIAL ORDERING OF MODELS

As described earlier $p_{ij}^{(r)} = g_r(\lambda_i/\lambda_j)$ is increasing in the ratio of scale parameters and, for fixed $\lambda_i < \lambda_j$, is decreasing in r. These facts are illustrated in Figure 1 which shows the value of $p_{ij}^{(r)}$ for r between 0.01 and 100 when $\lambda_i < \lambda_j$. The value of $p_{ij}^{(r)}$ for $\lambda_i > \lambda_j$ is obtained from $g_r(\lambda_i/\lambda_j) = 1 - g_r(\lambda_j/\lambda_i)$. As illustrated in Figure 1, different ratios of the λ_i are required to obtain the same value of p_{ij} for different values of r. It is ordinarily the case that the estimates of λ_i, $i = 1, \ldots, k$, which are denoted by $\hat{\lambda}_i$, $i = 1, \ldots, k$, are of less interest than the fitted preference probabilities $\hat{p}_{ij} = g_r(\hat{\lambda}_i/\hat{\lambda}_j)$. For example, in comparing the results of different paired comparison models, indexed by different values of r, we find the fitted values to be the relevant means of comparison.

A property of all linear models is that p_{ik} can be computed from p_{ij} and p_{jk}. The formula for this computation, called the triples function by Yellott (1977) and the composition rule by Latta (1979), defines a function $G(\cdot, \cdot)$ such that $p_{ik} = G(p_{ij}, p_{jk})$. For the gamma model with shape parameter r, the compositon rule can be expressed in terms of the inverse function $g_r^{-1}(p) = \{\gamma : g_r(\gamma) = p\}$, where γ is a ratio of scale parameters. The inverse is well defined since $g_r(\gamma)$ is monotone in γ. The composition rule for the gamma model with shape parameter r is

$$p_{ik}^{(r)} = G(p_{ij}^{(r)}, p_{jk}^{(r)}) = g_r\{g_r^{-1}(p_{ij}^{(r)})\ g_r^{-1}(p_{jk}^{(r)})\}. \tag{2}$$

As an illustration consider the Bradley-Terry model, where $g_r(\gamma) = \gamma/(\gamma + 1)$ and $g_r^{-1}(p) = p/(1-p)$. If $p_{ij}^{(1)} = 0.6$ and $p_{jk}^{(1)} = 0.8$, then $\lambda_i/\lambda_j = 1.5$ and $\lambda_j/\lambda_k = 4$, from which $\lambda_i/\lambda_k = 6$ and $p_{ik}^{(1)} = 0.857$. The composition rule is illustrated in Table

1 for a variety of values of p_{ij}, p_{jk} and r. Table 1 also includes the composition rule for convolution type linear models with the normal sensation distribution and the exponential sensation distribution. The values shown in Table 1 are for the section of the unit square in which $p_{jk} > 1/2, \quad (1 - p_{jk}) < p_{ij} \leq p_{jk}$. Values of the composition rule p_{ij}, p_{jk} in other sections of the unit square (e.g. $p_{ij} = 0.6, p_{jk} = 0.1$) can be obtained by applying the following properties of the composition rule G (Latta 1979):

(i) $G(p_{ij}, \frac{1}{2}) = p_{ij}$

(ii) $G(p_{ij}, 1 - p_{ij}) = \frac{1}{2}$

(iii) $G(p_{ij}, p_{jk}) = G(p_{jk}, p_{ij})$ (symmetry)

(iv) $G(p_{ij}, p_{jk}) = 1 - G(1 - p_{ij}, 1 - p_{jk})$

(v) $p_{ik} = G(p_{ij}, p_{jk}) \iff p_{ij} = G(p_{ik}, 1 - p_{jk}) \iff p_{jk} = G(1 - p_{ij}, p_{ik})$.

These properties are easy to verify for the gamma models. Consider property (iii) which is proved by a series of equalities using $g_r^{-1}(p) = 1/g_r^{-1}(1 - p)$ and $g_r(\gamma) = 1 - g_r(1/\gamma)$,

$$\begin{aligned} G(p_{ij}, p_{jk}) &= g_r\left(g_r^{-1}(p_{ij})\ g_r^{-1}(p_{jk})\right) = g_r\left(\frac{1}{g_r^{-1}(1 - p_{ij})}\ \frac{1}{g_r^{-1}(1 - p_{jk})}\right) \\ &= 1 - g_r\left(g_r^{-1}(1 - p_{ij})\ g_r^{-1}(1 - p_{jk})\right) \\ &= 1 - G(1 - p_{ij}, 1 - p_{jk}). \end{aligned}$$

Table 1 indicates that there is not much change in the value of p_{ik} obtained for fixed p_{ij} and p_{jk} as r varies from 0.01 to 100. The limiting behavior of the gamma models for small and large r is also demonstrated in Table 1. Burke and Zinnes (1965) found that the composition rules of the Bradley-Terry and Thurstone-Mosteller models are quite similar. This result is also demonstrated in Table 1.

Latta (1979) introduces a partial ordering on paired comparison models. The paired comparison model A is more extreme than the paired comparison model B if

for all $(p_{ij}, p_{jk}) \in \{(0.5, 1.0) \times (0.5, 1.0)\}$ $p_{ik}^{(A)} \geq p_{ik}^{(B)}$ with strict inequality for some pair. As before, the definition is given in terms of one quadrant of the unit square, since the definition is extended to the remainder of the unit square via properties (i)-(v) above. Latta gives an algebraic proof that the Thurstone-Mosteller model is more extreme than the Bradley-Terry model and proves the following theorem that gives a sufficient condition for determining whether one linear model is more extreme than a second in terms of the densities of the defining distributions.

Theorem (Latta 1979 p.369): Suppose that

(A) h_a and h_b are densities whose associated c.d.f.'s, H_a and H_b, satisfy the two conditions (1) $H(x) = 1 - H(-x)$ and (2) $H^{-1}(p)$ exists for $p \in (0, 1)$.

(B) for every $c > 0$ there exists $N_1(c) > N_2(c) \geq 0$ such that

(i) $|t| < N_2(c) \Rightarrow h_b(t) < ch_a(ct)$

(ii) $N_2(c) < |t| < N_1(c) \Rightarrow h_b(t) > ch_a(ct)$

(iii) $|t| > N_1(c) \Rightarrow h_b(t) < ch_a(ct)$.

Then the linear model based on h_b is more extreme than the linear model based on h_a.

The following proposition applies this theorem to gamma paired comparison models.

Proposition 1. If $r_1 > r_2$ then the gamma paired comparison model with shape parameter r_1 is more extreme than the gamma paired comparison model with shape parameter r_2.

Proof. The result is demonstrated by showing that the conditions in Latta's theorem are satisfied by the densities

$$h_{r_1}(x) = \frac{\Gamma(2r_1)}{\Gamma(r_1)\Gamma(r_1)} \frac{e^{-r_1 x}}{(1+e^{-x})^{2r_1}} \qquad \text{and} \qquad h_{r_2}(x) = \frac{\Gamma(2r_2)}{\Gamma(r_2)\Gamma(r_2)} \frac{e^{-r_2 x}}{(1+e^{-x})^{2r_2}}.$$

The densities satisfy the conditions in (A) and therefore we consider the ratio

$$R(t) = \frac{h_{r_1}(t)}{c h_{r_2}(ct)} = \frac{1}{c}\frac{\Gamma(2r_1)\Gamma(r_2)\Gamma(r_2)2^{2r_2}}{\Gamma(2r_2)\Gamma(r_1)\Gamma(r_1)2^{2r_1}}(\cosh\frac{t}{2})^{-2r_1}(\cosh\frac{ct}{2})^{2r_2},$$

where $\cosh(x) = (e^x + e^{-x})/2$, and the derivative of the ratio

$$\frac{\partial R}{\partial t} = R(t)(cr_2 \tanh\frac{ct}{2} - r_1 \tanh\frac{t}{2}),$$

where $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. As the densities h_{r_1}, h_{r_2} and the ratio $R(t)$ are symmetric we consider only $t \geq 0$. Form the function

$$k(r) = \frac{\Gamma(2r)}{\Gamma(r)\Gamma(r)2^{2r}}$$

from the coefficient of $h_r(x)$. Then it can be shown, using formulas for $\Gamma(r)$ and $\Gamma'(r)$ from Chapter 6 of Abramowitz and Stegun (1964), that $k(r)$ is increasing in r and $k(r)/r$ is decreasing in r. The conditions (B) of the theorem are verified by considering c in three regions, $c \leq 1$, $c \geq r_1/r_2$ and the intermediate range.

For $c \leq 1$, $\partial R/\partial t = 0$ for $t = 0$ and $\partial R/\partial t < 0$ for $t > 0$. Also, $R(0) > 1$ since $r_1 > r_2$, $c \leq 1$, and $k(r)$ is increasing in r, and $R(t)$ is less than 1 as $t \to \infty$. Thus, h_{r_1} starts above h_{r_2}, the densities cross once and then h_{r_1} remains below h_{r_2} after the crossing. The conditions of the theorem are satisfied with $N_2(c) = 0$. In a similar manner, we find that when $c \geq r_1/r_2$, $\partial R/\partial t > 0$ for $t \geq 0$, $R(0) < 1$ (since $k(r)/r$ is decreasing in r), and $R(t)$ is greater than 1 as $t \to \infty$. The conditions of the theorem are satisfied with $N_1(c) = \infty$.

For intermediate values of c, $R(0)$ may be greater than one, less than one or equal to one. However, the derivative has at most one change of sign, as can be verified by showing that the ratio $(cr_2 \tanh cx)/(r_1 \tanh x)$ is monotone decreasing. It turns out that for $c \leq \sqrt{r_1/r_2}$ there are no changes of sign of the derivative and for $c > \sqrt{r_1/r_2}$ the derivative is initially positive and becomes negative. If

$R(0) \leq 1$ then $R(t)$ increases initially and then decreases below one and remains there as $t \to \infty$, whereas if $R(0) > 1$ then $R(t)$ may decrease or increase initially but eventually ends below one. In either case, the conditions of the theorem hold, as the densities intersect at most twice (equivalently the ratio $R(t)$ is equal to one for at most two values of t). Thus, the conditions of the theorem are verified for all values of $c > 0$. •

This section and the preceding section focus attention on a subset of the convolution type linear models for paired comparisons experiments. The gamma paired comparison models include the most popular paired comparison models and are ordered by the extremeness of their composition rules. After briefly discussing inference for paired comparisons experiments, the empirical phenomenon that many models provide similar fits to a data set is examined by considering models that are extreme points in the family of gamma models.

4. INFERENCE

In the paired comparisons experiment with k objects, i and j are compared $n_{ij} = n_{ji}$ times, with i preferred to j in a_{ij} of the comparisons. No ties are permitted. If successive comparisons are independent, then a_{ij} is a binomial random variable with n_{ij} trials and the probability of a success on any trial is $g_r(\lambda_i/\lambda_j)$. Finally, if comparisons among different pairs of objects are independent then the likelihood for the entire data set is the product of $\binom{k}{2}$ binomial likelihoods. For fixed r, the maximum likelihood estimates of the scale parameters λ_i are obtained using a combination of Newton-Raphson and steepest descent steps. This approach works well except for small values of r, where an iterative approach (Ford 1957, Stern 1987) is required until the solution is nearby. The likelihood can not be maximized

if one object is always preferred to its competitors or if one object is never preferred to its competitors. To maximize the likelihood over r, the likelihood is evaluated for a grid of r values. This is more straightforward than directly incorporating r into the Newton-Raphson/steepest descent maximization.

To assess goodness of fit, consider the likelihood ratio test for the null hypothesis that the gamma model with shape parameter r (viewed as being fixed for the purposes of this discussion) is adequate versus the alternative hypothesis that maximizes each binomial likelihood separately. In the latter case, p_{ij} is estimated by a_{ij}/n_{ij}, while in the former p_{ij} is estimated by $g_r(\hat{\lambda}_i/\hat{\lambda}_j)$. The alternative hypothesis might be preferred if the data contains many inconsistent triads of the form $p_{ij} > 0.5$, $p_{jk} > 0.5$, $p_{ki} > 0.5$. These triads are not consistent with the property of strong stochastic transitivity ($p_{ij}, p_{jk} \geq 1/2$ implies $p_{ik} > \max(p_{ij}, p_{jk})$) (David 1988) that is implicitly assumed by all convolution type linear models. The usual test statistic for the above hypothesis, which we use as a measure of goodness of fit, is

$$Q_1 = 2\sum_{i=1}^{k}\sum_{j\neq i} a_{ij} \log \frac{a_{ij}/n_{ij}}{g_r(\hat{\lambda}_i/\hat{\lambda}_j)} .$$

If the gamma model is correct and the n_{ij} are large, then Q_1 has the chi-square distribution with the number of degrees of freedom equal to the difference between the number of free parameters in the two likelihoods, $\frac{1}{2}(k-1)k - (k-1) = \frac{1}{2}(k-1)(k-2)$. In practice, r is estimated and should be treated as a parameter for purposes of the goodness of fit test. However, models with different values of r are considered as different models in the following section and then compared to each other. Therefore r is treated as fixed in the next section. Notice that the usual likelihood ratio procedure can not be used to test whether one gamma model is superior to another since the models are not nested. Q_1 is used to compare the fit

of the models in the following section.

5. ARE ALL LINEAR MODELS THE SAME?

Consider the 1989 American League baseball season as a paired comparisons experiment to determine the relative ability of the fourteen teams. In the following matrix A, each team is represented by one row and column. The entries in row i, a_{ij}, correspond to the number of wins for team i in contests with team j.

$$A = \begin{pmatrix}
- & 6 & 8 & 7 & 6 & 8 & 11 & 5 & 5 & 5 & 7 & 3 & 7 & 11 \\
7 & - & 6 & 7 & 8 & 7 & 10 & 5 & 6 & 6 & 9 & 4 & 6 & 6 \\
5 & 7 & - & 6 & 7 & 8 & 11 & 7 & 4 & 4 & 6 & 6 & 5 & 7 \\
6 & 6 & 7 & - & 8 & 10 & 7 & 5 & 4 & 5 & 5 & 9 & 7 & 2 \\
7 & 5 & 6 & 5 & - & 4 & 7 & 3 & 6 & 6 & 5 & 6 & 8 & 6 \\
5 & 6 & 5 & 3 & 9 & - & 5 & 2 & 8 & 7 & 7 & 5 & 6 & 5 \\
2 & 3 & 2 & 6 & 6 & 8 & - & 4 & 6 & 1 & 4 & 5 & 4 & 8 \\
7 & 7 & 5 & 7 & 9 & 10 & 8 & - & 6 & 8 & 8 & 7 & 9 & 8 \\
7 & 6 & 8 & 8 & 6 & 4 & 6 & 7 & - & 9 & 8 & 7 & 9 & 7 \\
7 & 6 & 8 & 7 & 6 & 5 & 11 & 5 & 4 & - & 6 & 11 & 7 & 8 \\
5 & 3 & 6 & 7 & 7 & 5 & 8 & 5 & 5 & 7 & - & 8 & 7 & 10 \\
9 & 8 & 6 & 3 & 6 & 7 & 7 & 6 & 6 & 2 & 5 & - & 7 & 8 \\
5 & 6 & 7 & 5 & 4 & 6 & 8 & 4 & 4 & 6 & 6 & 6 & - & 6 \\
1 & 6 & 5 & 10 & 5 & 7 & 4 & 5 & 6 & 5 & 3 & 5 & 7 & -
\end{pmatrix}$$

The fourteen teams in the American League are divided into two seven team divisions. The top seven rows represent the teams in one division and the bottom seven rows represent the teams in the other division. Teams play 13 games against each team in their division and 12 games against each team in the other division. No ties are possible. One game, between team 5 and team 14, was cancelled due to inclement weather.

We consider the fit obtained by applying gamma paired comparison models to the results of baseball games even though the point scoring process in baseball is not similar to a Poisson process. The maximum likelihood estimates for gamma models with r ranging from 0.1 to 50 were obtained, and the goodness of fit statistic

Q_1 computed for each model. The values of Q_1 range from 81.47 for $r = 0.1$ to 81.19 for $r = 50$. The Bradley-Terry model has $Q_1 = 81.22$. The maximum of the likelihood, equivalent to the minimum value of Q_1, over the values of r considered here is obtained at $r = 50$ (approximately the Thurstone-Mosteller) model. On the one hand, we have the positive result that the gamma models provide an adequate fit to the data (values of Q_1 should be compared to the chi-square distribution with 78 degrees of freedom in this case). However, the variation among models is so small that no model is obviously preferred to the others. If r is viewed as a parameter of the model and $\hat{r}$ indicates the value of r that maximizes the likelihood, then an asymptotic 95% confidence interval for r includes all values of r such that Q_1 within 3.84 (the upper 5% point of the chi-square distribution on one degree of freedom) of the minimum value of Q_1. For this data set the confidence interval contains all values of r between 0 and 50 (larger r were not considered). The largest difference between the residuals of one gamma model (the difference between the matrix A and the fitted values obtained by a given model) and the residuals of a second is 0.17. The magnitude of the residuals range from 0 to 4.79 so that the variation among models is much smaller than one might expect. Large residuals typically correspond to extreme results, pairs in which one team dominates another despite the fact that each team won at least 35% of their games overall. The results for the 1989 American League season as well as nine other baseball data sets and five recent basketball seasons (teams play each other between 2 and 5 times) are given in Table 2. The results from five football seasons, in which teams play each other 0, 1 or 2 times, are also given. The chi-square approximation is inappropriate for the football data due to the small sample sizes. However, the similarity of the fit provided by different values of r is striking. In each case but one, the values of Q_1 are either monotone increasing or monotone decreasing indicating that the "best"

model is obtained by using the largest or smallest value of r. The results of the sports data sets reinforce the earlier results of Mosteller (1958) and Jackson and Fleckenstein (1957).

To investigate more thoroughly why this occurs, some calculations for artificial data are considered. Consider data that is generated from the gamma model with shape parameter $r = 0.1$, $p_{ij} = 0.9$, $p_{jk} = 0.9$, and as indicated by the composition rule, $p_{ik} = 0.9803$. Initially assume that 100 comparisons of each pair are carried out, with results exactly matching the model, i.e. i is preferred to j in 90 out of 100 comparisons, j is preferred to k in 90 out of 100 comparisons, and, to be precise, i is preferred to k in 98.03 comparisons. This represents a data set with no sampling variability. Gamma models with other values of r can be fit to this "observed" data, equivalent to misspecifying the model. Naturally, $r = 0.1$ provides a perfect fit to the data, $Q_1 = 0$. Even the most extreme model considered, $r = 50$, has a small value of the goodness of fit statistic $Q_1 = 1.58$. Recall that, for an experiment with 3 objects, when testing a particular gamma model against the alternative that each p_{ij} is estimated separately, Q_1 can be compared to the chi-square distribution on 1 degree of freedom. Thus 100 comparisons per pair are not sufficient to reject the $r = 50$ model when the data is generated by the $r = 0.1$ model with no error or variability. Noether (1960) applied the same approach using an alternative measure of fit. Using Q_1 enables us to determine the sample size required to distinguish between models. At usual significance levels, 250 observations of each pair are required to reject the $r = 50$ model as inadequate (compared to the saturated model) when the data is generated by the $r = 0.1$ model. The same analysis was repeated for a variety of p_{ij} and p_{jk} values, specifically, a grid where p_{ij} and p_{jk} were multiples of 0.05. The result described above is the scenario for which the models differed by the largest amount. In other cases 500, 1000 or more comparisons of

each pair are required to distinquish the $r = 0.1$ model from the $r = 50$ model. Even larger sample sizes are required to distinguish the Bradley-Terry model ($r = 1$) from other gamma models.

The previous analysis and that of Noether (1960) ignore the variability that occurs in samples. If random paired comparisons experiments are simulated in which 100 comparisons of each of the three pairs of objects, i versus j, i versus k, and j versus k, are generated, then the results are similar. For the example discussed in the preceding paragraph, the average goodness of fit statistic over 1000 replications for the model that generated the data ($r = 0.1$), was 1.133 and the standard deviation of the statistics was 1.534 (consistent with the null distribution, chi square on one degree of freedom). The average goodness of fit statistic for the $r = 50$ model is 2.619 and the standard deviation is 3.018. The average difference between the two models is 1.486, slightly smaller than the result obtained from data with no variability. The $r = 50$ model provides a better fit than the model that generated the data in 31% of the samples. Simulations for five objects indicate again that several hundred comparisons of each pair are required to distinguish between models. The required sample size is smallest in those data sets for which some of the p_{ij} are extreme.

For experiments with fewer comparisons of each pair, the extreme probabilities used above frequently produce simulated data sets such that i is always preferred to j and k. Maximum likelihood estimates can not be obtained for such data sets. Simulations were carried out using less extreme values of p_{ij}, p_{jk}, p_{ik}. Consider 1000 simulated data sets consisting of 20 comparisons of each pair of three objects with $r = 0.1$, $p_{ij} = 0.6$, $p_{jk} = 0.9$, $p_{ik} = 0.9210$. The average difference between the goodness of fit statistic for $r = 0.1$ and the goodness of fit statistic for $r = 50$ is 0.205. The incorrect model, $r = 50$, is preferred for 43% of the data sets. It is more

difficult to distinguish between the models in this case due to the decreased sample size (number of comparisons) and the less extreme preference probabilities.

6. DISCUSSION

The sports data sets and simulations seem to answer Mosteller's (1958 pg 284) call to "explore the sensitivity of the method of paired comparisons to the shape of the curve used to grade the responses". The gamma models provide a convenient family of models indexed by a single parameter that can be used to explore the question. By comparing models at extreme values of the shape parameter, the Thurstone-Mosteller model (r large) and the exponential model (r near zero), over a wide range of data sets and simulation scenarios, we find that the paired comparisons analysis is not very sensitive to the choice of distribution within the class of linear models. Moreover, in experiments with three objects, it appears that at least 250 comparisons of each pair of objects are required to distinguish between models using a goodness of fit test statistic. The work of Mosteller (1958) and Noether (1960) shows that the linear model defined by the uniform distribution (not part of the gamma models but more extreme than even the Thurstone-Mosteller model) also provides a similar fit.

In part, this result seems to be an example of the similarity of many distributions at the center of the distribution (see Cox 1970 for more details). The similarity between the fits obtained with the Bradley-Terry and Thurstone-Mosteller models is not surprising given the similarity of the logistic and normal distribution functions. The linearity assumption of the paired comparison models is also a part of the explanation. This assumption leads us to only consider strongly transitive models as the k objects are assumed to be rank ordered on a linear scale. The particular

distribution used to fit the linear model does not seem to be as important as the determination of whether a linear model is appropriate.

Some data sets will be consistent with simpler models, for example the objects may be organized as groups of similar objects. Then a linear model with some parameters set equal to each other will be sufficient. In other cases, those with inconsistencies for instance, a model that assigns one parameter per object will not be sufficient. This leads to more sophisticated models (Davidson and Bradley 1969, Hiyashi 1964, Marley 1988) that allow objects to be compared on one of several possible dimensions. Item i might be preferred to item j on one dimension but j might be preferred on other dimensions. The outcome of a paired comparison depends on which dimension(s) are used to compare the objects. The nature of the comparison experiment must dictate which model is appropriate. The comprehensive study here suggests that if a linear model is selected, the particular linear model does not have a large effect on the analysis for the usual sample sizes.

The similarity of fits among the linear models seems to also hold in experiments in which more than two objects are compared at a time. The order statistics ranking models described by Critchlow, Fligner and Verducci (1990) are the natural extension of the linear models to such experiments. Simulations like those described here indicate that the fit obtained by order statistics models is not sensitive to the distribution used.

ACKNOWLEDGEMENTS

The author thanks Herman Chernoff, Don Rubin and Andrew Gelman for help in various arguments. Two referees and the editor of this issue suggested many improvements and brought several important references to my attention. This work

was partially supported by National Science Foundation grant SES-8805433 and Office for Naval Research grant N00014-86K-0246.

REFERENCES

M. Abramowitz and I. A. Stegun, eds., Handbook of Mathematical Functions (National Bureau of Standards, Washington, 1964).

R. A. Bradley, Some statistical methods in taste testing, Biometrics 9 (1953) 22-38.

R. A. Bradley and M. E. Terry, Rank analysis of incomplete block designs. I. the method of paired comparisons, Biometrika 39 (1952) 324-345.

C. J. Burke and J. L. Zinnes, A paired comparison of paired comparisons, Journal of Mathematical Psychology 2 (1965) 53-76.

D. R. Cox, The Analysis of Binary Data (Chapman and Hall, London, 1970) pg 28.

D. E. Critchlow, M. A. Fligner and J. S. Verducci, Probability models on rankings, Journal of Mathematical Psychology 34 (1990) in press.

H. A. David, The Method of Paired Comparisons (second edition) (Griffin, London, 1988).

R. R. Davidson, On a relation between two representations of a model for paired comparisons, Biometrics 25 (1969) 597-599.

R. R. Davidson and R. A. Bradley, Multivariate paired comparisons: the extension of a univariate model and associated estimation and testing procedures, Biometrika 56 (1969) 81-95.

L. R. Ford Jr., Solution of a ranking problem from binary comparisons, American Mathematical Monthly 64 (1957) 28-33.

C. Hiyashi, Multidimensional quantification of the data obtained by the method of paired comparisons, Annals of the Institute of Statistical Mathematics 16

(1964) 231-245.

J. E. Jackson and M. Fleckenstein, An evaluation of some statistical techniques used in the analysis of paired comparisons, Biometrics 13 (1957) 51-64.

H. Joe, Majorization, entropy, and paired comparisons, Annals of Statistics 16 (1987) 915-925.

R. B. Latta, Composition rules for probabilities from paired comparisons, Annals of Statistics 7 (1979) 349-371.

R. D. Luce, Individual Choice Behavior (Wiley, New York, 1959).

R. D. Luce and P. Suppes, Preference, utility, and subjective probability, in: R. D. Luce, R. R. Bush, and E. Galanter, eds., Handbook of Mathematical Psychology, Vol. 3 (Wiley, New York, 1965).

A. A. J. Marley, A random utility family that includes many of the "classical" models and has closed form choice probabilities and choice reaction times, manuscript, Department of Psychology, McGill University (1988).

F. Mosteller, Remarks on the methods of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. II. the effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. III. a test of significance for paired comparisons when equal standard deviations and equal correlations are assumed, Psychometrika 16 (1951) 3-9, 203-206, 207-218.

G. E. Noether, Remarks about a paired comparison model, Psychometrika 25 (1960) 357-367.

J. H. Smith, Adjusting baseball standings for strength of teams played, American Statistician 10 (1956) 23-24.

H. Stern, Gamma processes, paired comparisons and ranking, Ph.D. Thesis, Department of Statistics, Stanford University, Stanford, CA (1987).

H. Stern, A continuum of paired comparisons models, Biometrika 77 (1990) 265-273.

L. L. Thurstone, A law of comparative judgment, Psychological Review 34 (1927) 273-286.

J. I. Yellott Jr., The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution, Journal of Mathematical Psychology 15 (1977) 109-144.

Figure 1

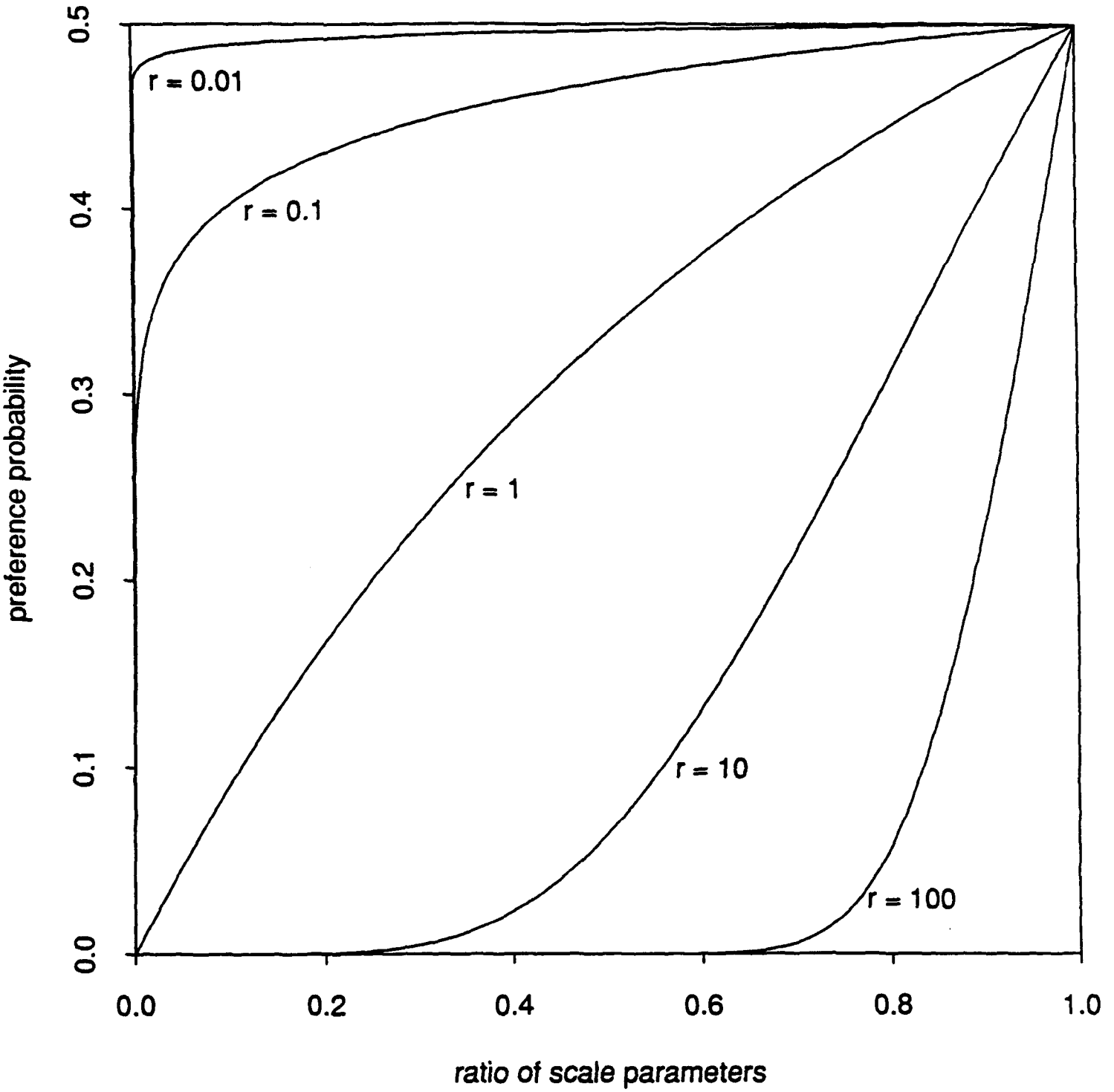
preference probability
0.0
0.1
0.2
0.3
0.4
0.5
r = 0.01
r = 0.1
r = 1
r = 10
r = 100
0.0
0.2
0.4
0.6
0.8
1.0
ratio of scale parameters

Figure 1. Preference probabilities in the gamma paired comparison model as a function of the ratio of the scale parameters λ_i/λ_j.

Table 1. Value of p_{ik} Obtained for Different Gamma Models

p_{ij}	p_{jk}	Exponential	$p_{ik}^{(0.01)}$	$p_{ik}^{(0.1)}$	$p_{ik}^{(1)}$	$p_{ik}^{(10)}$	$p_{ik}^{(100)}$	Normal
0.2	0.9	.75000	.74996	.74639	.69231	.67218	.67022	.67001
0.3	0.8	.66667	.66661	.66198	.63158	.62513	.62453	.62446
0.3	0.9	.83333	.83331	.83094	.79412	.77743	.77571	.77552
0.4	0.7	.62500	.62494	.62055	.60870	.60700	.60684	.60682
0.4	0.8	.75000	.74996	.74681	.72727	.72236	.72188	.72183
0.4	0.9	.87500	.87498	.87340	.85714	.84905	.84817	.84807
0.6	0.6	.68000	.68005	.68356	.69231	.69367	.69380	.69382
0.6	0.7	.76000	.76004	.76301	.77778	.78126	.78160	.78164
0.6	0.8	.84000	.84003	.84202	.85714	.86260	.86317	.86323
0.6	0.9	.92000	.92001	.92101	.93103	.93684	.93752	.93760
0.7	0.7	.82000	.82003	.82253	.84483	.85203	.85278	.85287
0.7	0.8	.88000	.88002	.88170	.90323	.91286	.91392	.91403
0.7	0.9	.94000	.94001	.94085	.95455	.96336	.96442	.96454
0.8	0.8	.92000	.92001	.92114	.94118	.95240	.95369	.95384
0.8	0.9	.96000	.96001	.96057	.97297	.98194	.98301	.98313
0.9	0.9	.98000	.98000	.98029	.98780	.99402	.99473	.99481

Table 2. Comparing Models on Sports Data Sets

League and Season	Teams	$r = 0.1$	$r = 1.0$	$r = 10.0$	$r = 50.0$
1989 American League Baseball	14	81.4704	81.2221	81.1914	81.1888
1986 American League Baseball	14	73.6317	73.6597	73.6610	73.6611
1985 American League Baseball	14	89.0463	89.1551	89.1786	89.1806
1984 American League Baseball	14	86.8979	86.8070	86.7829	86.7809
1983 American League Baseball	14	58.5468	58.5953	58.5932	58.5929
1989 National League Baseball	12	51.5812	51.5357	51.5314	51.5310
1986 National League Baseball	12	50.5396	50.0067	49.9121	49.9039
1985 National League Baseball	12	56.7934	56.6084	56.5811	56.5788
1984 National League Baseball	12	53.3042	53.4228	53.4357	53.4368
1983 National League Baseball	12	64.7119	64.7488	64.7516	64.7518
1981 National Basketball Assoc.	23	238.843	239.257	239.671	239.715
1980 National Basketball Assoc.	22	210.316	208.117	207.578	207.532
1979 National Basketball Assoc.	22	224.593	223.633	223.447	223.431
1978 National Basketball Assoc.	22	181.805	181.713	181.730	181.731
1977 National Basketball Assoc.	22	222.933	223.512	223.613	223.622
1986 National Football League	28	152.877	153.056	152.766	152.728
1985 National Football League	28	169.866	169.782	169.386	169.343
1984 National Football League	28	156.969	156.769	156.402	156.358
1983 National Football League	28	186.809	186.660	186.482	186.461
1981 National Football League	28	192.906	194.526	194.694	194.698

ABSTRACT

Previous authors (Jackson and Fleckenstein 1957, Mosteller 1958, Noether 1960) have found that different models of paired comparisons data lead to similar fits. This phenomenon is examined by means of a set of paired comparison models, based on gamma random variables, that includes the frequently applied Bradley-Terry and Thurstone-Mosteller models. A theoretical result provides a natural ordering of the models in the gamma family on the basis of their composition rules. Analysis of several sports data sets indicates that all of the paired comparison models in the family provide adequate, and almost identical, fits to the data. Simulations are used to further explore this result. Although not all approaches to paired comparisons experiments are covered by this discussion, the evidence is strong that for samples of the size usually encountered in practice all linear paired comparison models are virtually equivalent.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) TR No. ONR-C-5	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Dept. of Statistics Harvard University 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Department of Statistics SC713 Harvard University Cambridge, MA 02138	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING / SPONSORING ORGANIZATION ONR 8b. OFFICE SYMBOL (If applicable) Code 1111	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-91-J1005
8c. ADDRESS (City, State, and ZIP Code) Office of Naval Research Arlington, VA 22217-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)
Are All Linear Paired Comparison Models Equivalent?

12. PERSONAL AUTHOR(S)
Hal Stern

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	Sept., 1990	26

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Bradley-Terry model, Thurstone-Mosteller model

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

See reverse side.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL Herman Chernoff	22b TELEPHONE (Include Area Code) 617-549-5462 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED